AN UNBIASED CORRELATION RATIO MEASURE

BY TRUMAN L. KELLEY

HARVARD UNIVERSITY

Communicated July 16, 1935

The properties of the correlation ratio have been very thoroughly studied and reported upon. It has long been a necessary instrument in the study of the nature of regression. The work of Fisher[1] in 1922 made it a very precise instrument in studying the goodness of fit of second and higher degree regression lines.

It, however, lacks a certain desirable simplicity of meaning in that its value, η, obtained from a sample, differs from the population value, $\tilde{\eta}$, not only in a random manner due to the fluctuation of the particular sample, but also in a systematic manner which is a function of the number of arrays in which the data are recorded. This systematic difference between η and $\tilde{\eta}$ is, of course, well known to the expert statistician, and allowed for in his interpretations, as, for example, is automatically the case in the use of the following formula by Fisher

$$\chi^2 = (N - k)\frac{\eta^2 - R^2}{1 - \eta^2} \qquad (1)$$

where N is the number of cases in the sample, k the number of arrays in which the dependent variable is classed, η the ordinary correlation ratio, R the ratio of the standard deviation of the differences between the points upon the regression line used and the means of the arrays of the sample to the standard deviation of the dependent variable (for a linear regression line, R is simply r, the ordinary product-moment correlation coefficient), and χ^2 is the ordinary χ^2 distributed nearly in the Pearson type III manner and with a number of degrees of freedom equal to $[k - f(R)]$ in which $f(R)$ is the number of linear restrictions placed upon the frequencies in determining the regression line employed, it equaling 2 in the case in straight line regression, 3 for second degree parabolic regression, etc.

Entering a table giving probabilities for values of χ^2 with the value given by [1] and a number of degrees of freedom equal to $[k - f(R)]$ yields a value P which is the probability that if the true regression, or regression in the population, is of the form assumed, a divergence from it as great as that observed would arise as a matter of chance. Thus P, derived for χ^2, is an immediately interpretable statistic. We may note the simplicity of several of the other concepts.

N = the number of cases in the sample
k = the number of arrays

$k - f(R)$ = the number of degrees of freedom in the differences between the means of the arrays and the points on the regression line

R = an unbiased estimate of the R, which we will designate $\tilde{R}$, in the population.

However, η is not an unbiased estimate of the correlation ratio, $\tilde{\eta}$, in the population.

The intent of the present article is to provide a new correlation ratio squared, ϵ^2, which is an unbiased estimate of $\tilde{\eta}^2$; to provide its standard error; and to substitute it in place of η^2 in Fisher's formula [1]. So far as P and χ^2 are concerned, its use leads to identical results with [1].

The need to allow for the systematic effect of fine grouping for small numbers in arrays was handled by Pearson[2] in 1923, but he did not use Fisher's results of 1922. In 1906 Blakeman[3] provided a test for linearity which now, upon the basis of the later work of Pearson and Fisher, we must believe to be inadequate.

In the following treatment the notation will indicate population statistics by employing a tilda circumflex, ~; statistics which are means from a large number of samples by the caret, ^; statistics which are sample means by the macron, –; and other sample statistics by symbols having none of these circumflexes. The difference between the ^ and the ~ statistics is entirely a matter of the size of the samples entering into the average yielding the ^ statistics. With large samples every ^ statistic approaches the corresponding ~ statistic. Subscripts a, b, c, $\ldots k$ will refer to the k successive classes of the independent variable. Also subscript q will refer to any one such array. Letting v without subscript represent the variance of the dependent variable for the marginal total, and v with subscript the variance for an array, the true, or population, value of the correlation ratio is given by

$$\tilde{\eta}^2 = 1 - \frac{S\tilde{n}_a\tilde{v}_a}{\tilde{v}S\tilde{n}_a} \qquad (2)$$

wherein $\tilde{n}_a$ is the number giving the same proportion in array a as in the population, and S indicates a summation for all arrays, a, b, $\ldots$ k.

The usual or "raw" correlation ratio squared is given by the same formula dropping the circumflex. This value, η^2, is subject to a fineness of grouping error.

If the numbers in the arrays for a sample are n_a, n_b, $\ldots$ n_k giving proportions differing in a random manner from the proportions in the population, which are $\tilde{n}_a$, $\tilde{n}_b$, $\ldots$ $\tilde{n}_k$, the function,

$$\frac{Sn_a\tilde{v}_a}{Sn_a}$$

will differ by chance only from

$$\frac{S\tilde{n}_a\tilde{v}_a}{S\tilde{n}_a}.$$

Thus we may write the approximate equality

$$\tilde{\eta}^2 = 1 - \frac{Sn_a\tilde{v}_a}{\tilde{v}Sn_a}. \tag{3}$$

This is an estimate of $\tilde{\eta}^2$ having no systematic error. We now desire estimates of $\tilde{v}_a$, $\tilde{v}_b$, ... $\tilde{v}_k$, and $\tilde{v}$ having no systematic errors. As shown by "Student,"[4] these are readily available. They are, respectively,

$$\frac{n_a v_a}{n_a - 1}, \frac{n_b v_b}{n_b - 1}, \cdots \frac{n_k v_k}{n_k - 1}, \frac{Nv}{N - 1}$$

wherein $N = Sn_a$.

If we introduce these values in (3) we will obtain a formula for η^2 in which the estimated true variances for the arrays are weighted according to the number of cases in the sample in each array. Each array variance, i.e., each $\frac{n_c\ v_a}{n_a - 1}$, is an estimate of the residual variance in the dependent variable knowing the value (category) of the independent variable. However, the average of independent measures of the same thing having the least standard error is given by weighting each inversely as its variance, as early shown by Gauss.[5] The variance of

$$\frac{n_a\ v_a}{n_a - 1} \tag{4}$$

equals $\frac{n_a^2}{(n_a - 1)^2}$ times the variance of v_a, which for any distribution is as given by "Student"

$$v(v_a) = 2\tilde{v}_a^2 \frac{n_a - 1}{n_a^2} \tag{5}$$

We accordingly obtain the variance of (4)

$$\frac{2\tilde{v}_a^2}{n_a - 1}$$

Introducing the reciprocals of these as weighting factors in place of n_a, n_b, ... n_k in (3) yields

$$\epsilon^2 = 1 - \frac{S\left(\frac{n_a - 1}{2\tilde{v}_a}\right)\left(\frac{n_a\ v_a}{n_a - 1}\right)}{v\frac{N}{N - 1}\ S\frac{n_a - 1}{2\tilde{v}_a}} \tag{6}$$

It is entirely conceivable that $\bar{v}_a \neq \bar{v}_b \neq \bar{v}_c$, etc., but since these are of the nature of residual variances, it would not seem violent to assume them equal, in which case (6) becomes

$$\epsilon^2 = 1 - \frac{(N-1)Sn_a v_a}{(N-k)Nv}. \tag{7}$$

Since $\eta^2 = 1 - \frac{Sn_a v_a}{Nv}$, we may write (7) as

$$\epsilon^2 = \frac{(N-1)\eta^2 + 1 - k}{N-k} = \frac{N\eta^2 - k + (1-\eta^2)}{N-k}. \tag{8}$$

Comparing results here obtained with those of Pearson in 1923[6] we find certain small but significant differences.

Pearson gives the chance η^2 value, in case true $\eta^2 = 0$, as $\frac{(k-1)}{N}$, whereas from (8) we obtain $\frac{k-1}{N-1}$, which value has also been obtained by Wishart.[7] Pearson gives η^2 corrected for fineness of grouping, in case η^2 is fairly large, as equal to $\frac{N\eta^2 - k + 3}{N-k+3}$, which differs slightly from (8).

Let us now determine the standard error of ϵ^2, given by (7).

$$v(\epsilon^2) = \left(\frac{N-1}{N-k}\right)^2 v\left(\frac{Sn_a v_a}{Nv}\right) \tag{9}$$

$$Nv = Sx^2$$

in which x is the dependent variable as a deviation from the sample mean. For array a each x may be written

$$x_a = \bar{x}_a + (x_a - \bar{x}_a)$$

where $\bar{x}_a$ is the mean for the array as a deviation from the sample mean, and, as shown by "Student," $\bar{x}_a$ is uncorrelated with $v(x_a - \bar{x}_a)$. We may therefore write

$$Sx^2 = Sn_a\bar{x}_a^2 + Sn_a v_a \tag{10}$$

Let

$$f = \frac{Sn_a v_a}{Nv}. \tag{11}$$

Then

$$v(\epsilon^2) = \left(\frac{N-1}{N-k}\right)^2 v_f \tag{12}$$

And

$$v_\epsilon = \frac{v(\epsilon^2)}{4\epsilon^2} = \frac{(N-1)^2}{4(N-k)^2\epsilon^2} v_f \tag{13}$$

$$f = (1 - \epsilon^2)\frac{N-k}{N-1}$$

$$= \frac{Sn_a v_a}{Sn_a v_a + Sn_a \bar{x}_a^2} \tag{14}$$

$$\frac{df}{\hat{f}} = \frac{dSn_a v_a}{Sn_a v_a} - \frac{dSn_a v_a + dSn_a \bar{x}_a^2}{Sn_a \hat{v}_a + Sn_a \bar{x}_a^2} \tag{15}$$

Employing the earlier assumption that $\tilde{v}_a = \tilde{v}_b = \ldots \tilde{v}_k$

$$Sn_a \hat{v}_a = Sn_a \frac{n_a - 1}{n_a}\tilde{v}_a = (N-k)\tilde{v}_a \tag{16}$$

$$S\hat{x}^2 = (N-1)\tilde{v}. \tag{17}$$

Taking logarithmic differentials of (14)

$$\frac{df}{\hat{f}} = \frac{dSn_a v_a}{(N-k)\tilde{v}_a} - \frac{dNv}{(N-1)\tilde{v}}. \tag{18}$$

Squaring, summing and dividing by the number of samples

$$\frac{vf}{\hat{f}^2} = \frac{S2n_a^2\tilde{v}_a^2\dfrac{n_a-1}{n_a^2}}{(N-k)^2\tilde{v}_a^2} + \frac{2N^2\tilde{v}^2\dfrac{N-1}{N^2}}{(N-1)^{2\sim 2}} - \frac{2S2n_a^2\tilde{v}_a^2\dfrac{n_a-1}{n_a^2}}{(N-k)(N-1)\tilde{v}_a\tilde{v}}$$

$$= \frac{2}{N-k} + \frac{2}{N-1} - \frac{4(1-\eta^2)}{N-1}. \tag{19}$$

From (12)

$$v(\epsilon^2) = \frac{(1-\epsilon^2)^2(N-k)^2(N-1)^2}{(N-1)^2(N-k)^2}\left\{\frac{2}{N-k} - \frac{2-4\epsilon^2}{N-1}\right\}$$

$$v(\epsilon^2) = \frac{(1-\epsilon^2)^2}{N-1}\left\{\frac{2(k-1)}{N-k} + 4\epsilon^2\right\} \tag{20}$$

$$\sigma_{\epsilon^2} = \frac{1-\epsilon^2}{\sqrt{N-1}}\left\{\frac{2(k-1)}{N-k} + 4\epsilon^2\right\}^{1/2} \tag{21}$$

$$v_\epsilon = \frac{v(\epsilon^2)}{4\epsilon^2} = \frac{(1-\epsilon^2)^2}{4(N-1)\epsilon^2}\left\{\frac{2(k-1)}{N-k} + 4\epsilon^2\right\}\ldots\begin{Bmatrix}\epsilon^2 \text{ not}\\ \text{small}\end{Bmatrix} \tag{22}$$

$$\sigma_\epsilon = \frac{1-\epsilon^2}{2\epsilon\sqrt{N-1}}\left\{\frac{2(k-1)}{N-k} + 4\epsilon^2\right\}^{1/2}\ldots\begin{Bmatrix}\epsilon^2 \text{ not}\\ \text{small}\end{Bmatrix} \tag{23}$$

Formulas (22) and (23) are not usable formulas in case ϵ is small, as certain higher order terms have been neglected. Formulas (20) and (21) are

generally useful except when $\frac{1}{N}$ is not small in comparison with $\frac{1}{\sqrt{N}}$.

To obtain a test for goodness of fit we will make a substitution in (1). In place of η^2 we will substitute ϵ^2 as given by (8). This yields

$$\chi^2 = \frac{(N - k)(\epsilon^2 - R^2) + (k - 1)(1 - R^2)}{1 - \epsilon^2} \qquad (24)$$

$\{k - f(R)$ degrees of freedom$\}$

A second substitution is called for in that R^2 will differ in a slightly systematic manner from $\tilde{R}^2$. We have

$$R = \hat{R} + \delta R$$

and

$$\hat{R}^2 = \tilde{R}^2 + v(R).$$

If we define ρ^2 by the equation

$$R^2 = \rho^2 + v(R)$$

the quantity $(\epsilon^2 - \rho^2)$ and not $(\epsilon^2 - R^2)$ is a quantity whose mean value would equal zero in the case of correctness of the assumption as to the nature of the regression. However, the difference between $(\epsilon^2 - \rho^2)$ and $(\epsilon^2 - R^2)$ will ordinarily be so small that unappreciable gain in interpretability of the elements entering into (24) would result from incorporating ρ^2 instead of R^2 into the formula. As (24) stands, the magnitude of $(\epsilon^2 - R^2)$ is itself a good indication of goodness of fit. In particular, if $(\epsilon^2 - R^2) < 0$, an excellent fit is indicated without calculating χ^2 or P. If $(\epsilon^2 - R^2)$ is much less than zero, one should be skeptical of the arithmetical accuracy of his computations or of the logical soundness of some step in his treatment.

[1] R. A. Fisher, "The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients," *Jour. Roy. Stat. Soc.*, **85**, Part IV, July (1922).

[2] "On the Correction Necessary for the Correlation Ratio," *Biometrika*, **14** (1923).

[3] John Blakeman, "On Tests for Linearity of Regression in Frequency Distributions," *Ibid.*, **4** (1905–06).

[4] "The Probable Error of a Mean," *Ibid.*, **6**, No. 1 (1908).

[5] See Kelley, Truman L., *Statistical Method*, 1923, Formula [309].

[6] "On the Correction Necessary for the Correlation Ratio," *Biometrika*, **14** (1923).

[7] John Wishart, "A Note on the Distribution of the Correlation Ratio," *Ibid.*, **24** (1932).